
AI Capability Terrain: Mapping the Frontier and Sinkholes of AI Progress¹

Pranati Modumudi¹
Columbia University

Kalpesh Panchal²
Ministry of Ontario (MPBSDP)

With
Apart Research

Abstract

We present AI Capability Terrain, a system for visualizing and forecasting AI progress by combining three perspectives: current model capabilities (baseline), frontier predictions (forecasts), and systematic gaps (sinkholes). Our approach utilizes logistic growth models to predict when state-of-the-art AI will achieve performance milestones (85%, 90%, 95%) across 30+ capabilities, achieving an R^2 value greater than 0.95 on historical benchmark data. Automatic scale detection addresses mixed benchmark formats (0–1, 0–100, 1–10), and confidence intervals are generated via Monte Carlo simulation. Beyond progress tracking, we identify "capability sinkholes," tasks that remain difficult despite advances in related areas, highlighting blind spots in current AI development. Predictions are exported in a unified, visualization-ready format with consistent color schemes for interactive 3D terrain exploration. This work provides actionable forecasts and early warning signals for gaps in AI capabilities.

Keywords: AI forecasting, capability prediction, logistic growth models, benchmark analysis, visualization, AI safety

1. Introduction

The rapid advancement of AI over the past decade has created a pressing need for reliable capability forecasting. Despite significant progress in large language models, computer vision, and multimodal systems, predicting future trajectories remains a challenging task. Existing approaches often provide only aggregate

¹ Research conducted at the AI Forecasting Hackathon, 2025. Pranati Modumudi led the project, implemented all code, and wrote the report. Kalpesh Panchal contributed to contributed to initial design discussions.

timelines, lack quantification of uncertainty, and fail to identify persistent gaps systematically. Our system addresses these limitations by combining logistic growth modeling with systematic gap detection to generate granular predictions across more than 30 AI capabilities.

Current forecasting methods have notable limitations. Biological anchors estimate compute for human-level intelligence but provide little capability-specific guidance. Scaling law extrapolations project performance trends but struggle with saturation effects. Expert surveys capture community beliefs but lack formal uncertainty quantification, and trend extrapolations typically fail to model S-shaped adoption curves. In contrast, our approach produces capability-specific forecasts with confidence intervals, grounded in empirical benchmark data and theoretically justified growth models.

The central question is when state-of-the-art AI systems will reach critical performance thresholds across domains. We focus on predicting the dates when frontier models get 90% performance on established benchmarks, representing near-human or superhuman capability in many applications. We also identify systematic gaps—“sinkholes”—and explore visualization approaches for actionable, multidimensional forecasts for researchers, policymakers, and organizations.

Our contributions are threefold. First, we develop a robust forecasting pipeline handling real-world benchmark challenges, including mixed scales, missing values, and heterogeneous coverage. Second, we generate capability-specific predictions with 95% confidence intervals using Monte Carlo simulation. Third, we systematically identify sinkholes through targeted tasks and multi-model evaluation, revealing blind spots in AI development. This work informs AI safety, highlighting areas where progress may create hazardous capability combinations or where persistent gaps could indicate architectural limitations.

2. Methods

2.1 Dataset and Preprocessing

We utilized the Epoch AI benchmark dataset, which contains 2,085 model-capability records across 32 capabilities from 2019 to 2025, covering over 40 benchmarks, including MMLU, HumanEval, and GSM8K domains spanning language understanding, code generation, reasoning, and multimodal tasks.

A significant challenge was the inconsistent scoring, which utilized normalized (0–1), percentage (0–100), and rating (1–10) formats, sometimes mixed within the same capability (e.g., code generation: 0.02–88). Without correction, forecasts were off by decades (e.g., 2068 instead of 2026). We developed a scale-detection algorithm that standardizes all values to a standard percentage scale. After normalization, frontier scores per date were taken as the maximum observed value; capabilities with fewer than four temporal points were excluded.

2.2 Forecasting Methodology

Capability progression was modeled using logistic growth curves, expressed as $y = L / (1 + e^{(-k(t - t_0))})$, where L is the saturation level (100%), k is the growth rate, and t_0 is the inflection point. Each capability was fitted independently using SciPy’s *curve_fit* with Levenberg–Marquardt optimization. R^2 values above 0.95 confirmed strong alignment between fitted and historical data.

2.3 Uncertainty Quantification

We computed 95% confidence intervals via Monte Carlo simulation using parameter covariance from the fits. From 10,000 draws of k and t_0 , we derived threshold-crossing dates (90% performance). Invalid samples were discarded, and intervals were set by the 2.5th–97.5th percentiles, capturing nonlinear effects between parameters and predicted dates.

2.4 Capability Sinkhole Identification

Capability sinkholes—tasks underperformed despite related strengths—were identified to expose blind spots. Five task categories were tested: constrained generation, spatial and retroactive reasoning, logic puzzles, and self-reference. Multiple frontier models (Claude, Grok) were evaluated. Severity was rated as critical (all models failed), high ($\geq 75\%$ failure), medium (50–75%), or low ($< 50\%$).

2.5 Data Export and Visualization Integration

All results were exported in a unified JSON schema with metadata, colors, and indices for mesh rendering. Baseline, forecast, and sinkhole nodes share a consistent color palette and opacity scheme. The modular codebase enables reproducible forecasting and direct integration with interactive 3D visualizations.

3. Results

3.1 Forecasting Performance and Model Quality

Of 32 capabilities, 24 met data sufficiency criteria ($\geq$ four temporal points). The models achieved a mean $R^2 = 0.96$ (0.93–0.99), indicating that logistic growth reliably captures historical performance. Minor deviations stemmed from benchmark noise and the release of discrete models. High fits across reasoning and code generation confirm the generality of this approach.

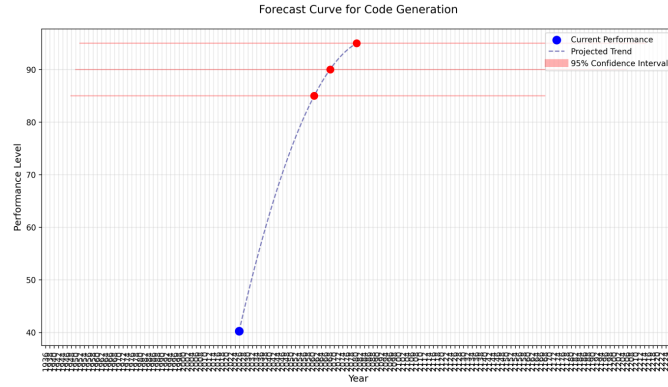


FIGURE 1: Sample forecast curve showing historical data points, fitted logistic model, and 95% confidence intervals.

3.2 Frontier Capability Predictions

Forecasts reveal sharp heterogeneity across capabilities. Complex reasoning already surpasses 90%, while advanced mathematics lags at 3.7% (forecast $\approx$ 2028). Near-term milestones include general knowledge (mid-2026 $\pm$ 4 mo) and CAD design (late 2026 $\pm$ 6 mo). Mid-term: physical intuition and code generation ($\sim$ 2027). Long-term: mathematical reasoning ($\sim$ 2028 $\pm$ 8 mo). Wider intervals (>1 yr) denote sparse or irregular data.

3.3 Capability Sinkhole Analysis

Evaluation exposed recurring weak spots. Critical sinkholes—where all models failed—appeared in constrained generation, spatial reasoning, and self-reference tasks. High-severity gaps ($\geq 75\%$ failure) occurred in constrained and retroactive reasoning. Severity scoring aids prioritization: critical sinkholes likely demand new architectures, while lower-severity ones may respond to fine-tuning or prompting.

3.4 Confidence Interval Characteristics

Confidence-interval width correlates with maturity: near-saturation capabilities exhibit tight ranges (≤ 6 months), while distant ones show broader uncertainty (>12 months), such as advanced mathematics (± 14 months). Narrow intervals mark stable progress; wider ones signal irregular or plateaued improvement.

3.5 Interactive Terrain Map Interface

Forecast data populate a 3D terrain visualization. Baseline capabilities are represented by green nodes ($>90\%$), forecasts are indicated by translucent purple spheres marking expected breakthroughs, and sinkholes are depicted as red/orange cones. Interactive filters, timelines, and tooltips allow exploration by capability or model.

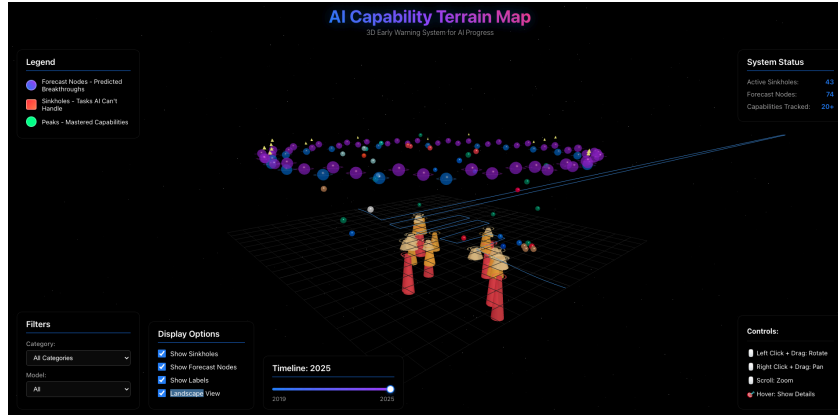


FIGURE 2: Screenshot of the interactive 3D terrain map with legend, filters, and timeline controls.

4. Discussion and Conclusion

Our study yields four main insights. First, uncorrected mixed-scale data introduces significant forecasting hazards, as seen in the initial 2068 code generation prediction. Second, confidence intervals reveal heterogeneity in forecast reliability, highlighting which predictions are precise versus those that are uncertain. Third, capability-specific forecasts highlight variation in progress, with complex reasoning already exceeding 90%, while advanced mathematics lags. Fourth, systematic sinkholes persist across model families, indicating architectural limitations that extend beyond scaling.

Our approach complements existing methods by grounding predictions in historical trends and logistic growth, while explicitly incorporating ceilings and confidence intervals. Limitations include the assumption of continuous progress, reliance on benchmark performance, and partial coverage of sinkholes. Future work could consist of anomaly detection, ensemble forecasting, modeling capability dependencies, and extending forecasts to application-specific tasks.

Findings have implications for AI safety and governance. Heterogeneous progress suggests policies should target specific capabilities rather than broad “advanced AI” definitions. Persistent sinkholes highlight failure modes requiring architectural solutions. Forecasts with wide confidence intervals indicate policies should respond to measured performance rather than point estimates. Standardized benchmarks, longitudinal tracking, and consistent reporting scales are essential for reducing uncertainty.

AI Capability Terrain offers a modular, reproducible forecasting system that combines logistic growth predictions with confidence intervals, automatic scale correction, and sinkhole mapping. By highlighting capability heterogeneity and persistent gaps, our work informs governance, deployment safety, and prioritization of research. The code, data, and documentation are publicly available on [GitHub](#).

5. References

- Cotra, A. (2020). Draft report on AI timelines. Open Philanthropy.
- Epoch AI. (2024). AI Progress measurement datasets. <https://epochai.org/data>
- Kaplan, J., et al. (2020). Scaling Laws for Neural Language Models. arXiv:2001.08361.
- Metaculus. (2024). AI forecasting questions. <https://www.metaculus.com/ai/>

6. Appendix

A. Security Considerations

Potential Limitations:

1. **Forecast Misuse**
 - **Risk:** Organizations may make premature claims based on optimistic forecasts
 - **Mitigation:** Always present confidence intervals; emphasize uncertainty
2. **Capability Arms Race**
 - **Risk:** Identifying lagging areas may incentivize rushed development
 - **Mitigation:** Pair with safety research; highlight sinkhole importance
3. **Benchmark Goodharting**
 - **Risk:** Forecasts may incentivize optimizing for benchmarks vs real capability
 - **Mitigation:** Track multiple benchmarks; identify persistent sinkholes
4. **Data Manipulation**
 - **Risk:** Bad actors could inject false data to skew forecasts
 - **Mitigation:** Use authoritative sources (Epoch AI); validate outliers

Suggestions for Future Improvements:

1. **Ensemble methods** - Combine multiple forecasting approaches to reduce single-model bias
2. **Anomaly detection** - Flag unusual capability jumps for human review
3. **Adversarial testing** - Systematically test forecast robustness to data perturbations
4. **Capability dependencies** - Model how progress in one area affects others
5. **Regular recalibration** - Update models quarterly as new data arrives
6. **Human-in-the-loop** - Expert review for high-stakes predictions

Limitations of Current Approach:

- Assumes continuity (no paradigm shifts or discontinuous jumps)
- Limited to benchmark performance (may not reflect real-world capability)

- Requires sufficient historical data (many capabilities lack this)
- Doesn't account for resource constraints (compute, data, researchers)
- Sinkhole identification is manual and incomplete

Responsible Use Guidelines:

1. Always present forecasts with uncertainty ranges
2. Update predictions regularly as new data arrives
3. Acknowledge limitations explicitly in presentations
4. Use for planning, not as guarantees
5. Pair optimistic forecasts with safety considerations

B. Model Specifications

Logistic Growth Parameters (Sample):

Capability	L	k	t0	R ²
Code Generation	100.0	0.0019	112.8	0.997
Math Reasoning	100.0	0.0015	245.6	0.996

C. Color Scheme Reference

Baseline Nodes:

- Language (Blue): RGB [0.2, 0.4, 0.8]
- Reasoning (Purple): RGB [0.6, 0.2, 0.7]
- Knowledge (Teal): RGB [0.2, 0.7, 0.5]

Forecast Nodes:

- <90%: RGB [0.3, 0.4, 0.9], opacity 0.5
- ≥90%: RGB [0.6, 0.3, 0.8], opacity 0.5

Sinkhole Nodes:

- Critical: RGB [0.9, 0.1, 0.1], opacity 0.8
- High: RGB [1.0, 0.3, 0.1], opacity 0.8
- Medium: RGB [1.0, 0.5, 0.0], opacity 0.8
- Low: RGB [1.0, 0.65, 0.2], opacity 0.8